A simple one dimensional glassy Kac model

Andrea Montanari

Departments of Electrical Engineering and Statistics,

Stanford University, Stanford CA-9305 USA

Antoine Sinton

Laboratoire de Physique Théorique de l'Ecole Normale Supérieure,

24 rue Lhomond 75231 Paris Cedex 05, France

We define a new family of random spin models with one-dimensional structure, finite-range multi-spin interactions, and bounded average degree (number of interactions in which each spin participates). Unfrustrated ground states can be described as solutions of a sparse, band diagonal linear system, thus allowing for efficient numerical analysis.

In the limit of infinite interaction range, we recover the so-called XORSAT (diluted p-spin) model, that is known to undergo a random first order phase transition as the average degree is increased. Here we investigate the most important consequences of a large but finite interaction range: (*i*) Fluctuation-induced corrections to thermodynamic quantities; (*ii*) The need of an inhomogeneous (position dependent) order parameter; (*iii*) The emergence of a finite mosaic length scale. In particular, we study the correlation length divergence at the (mean-field) glass transition.

PACS numbers: 64.70.Pf (Glass transitions), 75.10.Nr (Spin-glass and other random models), 89.20.Ff (Computer science)

I. INTRODUCTION

A large class of disordered mean field spin models exhibit a behavior that is reminiscent of the structural glass transition in fragile glasses [1, 2, 3]. As temperature is lowered, they undergo a *'dynamical phase transition'* characterized by a diverging relaxation time at a critical temperature $T_{\rm d}$. The reason for such a dynamical arrest can in turn be ascribed to ergodicity breaking: below $T_{\rm d}$ the Boltzmann measure decomposes into an exponential number of pure states. While equilibration is fast within each state, it takes an exponentially large (in the system size) time to equilibrate across states.

Below $T_{\rm d}$, the system can be meaningfully characterized through its *complexity* Σ, which gives the exponential growth rate of the number of pure states (i.e. the number of such states is about $e^{N\Sigma}$, N being the size). The complexity decreases as temperature is further lowered, and vanishes

linearly at a second (static) transition temperature $T_{\rm s}$. This corresponds to an actual thermodynamic phase transition.

A strikingly similar scenario has been found to hold in a large array of *random constraint satisfaction* [4, 5, 6] problems of interest in theoretical computer science[1]. The role of temperature is played here by the number of constraints per variable γ, while Boltzmann distribution is replaced by the uniform measure over solutions of the problems. As the constraint density crosses a critical value $\gamma_{\rm d}$, the set of solutions splits into 'lumps' analogous to pure states. Above a second threshold $\gamma_{\rm s}$ the set of constraints becomes with high probability unsatisfiable.

In the last few years there has been a consistent effort in interpreting disordered mean field models as a genuine mean field theory for the structural glass transition. This is highly non-trivial since in any finite-dimensional model there cannot be coexistence of an exponentially large number of pure states. Imagine trying to select one such state through appropriate boundary conditions on a box of size ℓ. This will imply an energetic bias towards the selected state, which is of order $\sigma\ell^\theta$, where $0 \le \theta \le d-1$ is a surface tension exponent. On the other hand, the entropic advantage of the other states is of order $\Sigma\ell^d$, because of their number. Therefore, for $\ell \gtrsim \ell_{\rm s} \equiv (\beta\sigma/\Sigma)^{\frac{1}{d-\theta}}$, pure states are no longer stable.

According to the *'mosaic state'* scenario, below $T_{\rm d}$ a typical configuration of the system can be described as a patchwork [7, 8, 9]. Each patch corresponds to the configuration being close to a particular pure state in a localized region whose length scale is $\ell_{\rm s}$. Since $\Sigma \sim (T - T_{\rm s})$ at the static transition, the mosaic lengthscale diverges as $\ell_{\rm s} \sim (T - T_{\rm s})^{-\nu}$ with $\nu = 1/(d-\theta)$.

While the mosaic scenario is appealing, its consistency and implications, as well as its precise meaning, are far from obvious. An important step forward was achieved in [10] where a concrete "gedanken experiment" was introduced to define $\ell_{\rm s}$. This length scale was interpreted in [11] as a *point-to-set* correlation length, and its divergence was rigorously proved to be equivalent to a divergent correlation time. In [12], $\ell_{\rm s}$ was actually shown to diverge at $T_{\rm s}$ in a class of disordered Kac models with continuous scalar spins.

Unhappily the models considered in [12] can currently be treated only in the Kac limit, and through somewhat formal techniques such as the replica method. As a consequence, many interesting questions (such as the relevance of this limit for realistic interaction ranges, non-perturbative fluctuation effects, a precise definition of states) cannot be addressed in this context. The present

[1] In a typical constraint satisfaction problems, one seeks an assignment of values to L discrete variables in such a way to satisfy M constraints.

paper aims at introducing a new class of models that share some features with the ones treated in [12], while being tractable within alternative approaches (e.g. numerically).

We follow the route of generalizing one of the ensembles of random constraint satisfaction problems mentioned above, and referred to as k-XORSAT [13, 14]. We shall require constraints to have finite range with respect to an underlying one-dimensional geometry. Our motivation is twofold: (*i*) Because of its underlying linear structure, the k-XORSAT is very well understood. In particular a wealth of informations regarding pure states and their geometry is accessible through rigorous techniques [15, 16, 17, 18]; (*ii*) The ensembles of random constraint satisfaction problems studied within the computer science and statistical mechanics communities have lacked so far any geometrical structure (in physics terms, they are mean field models). This is of course a poor cartoon of real world instances, and it is surely instructive to explore alternative –structured– models.

Constraint satisfaction problems with finite interaction range were already considered in [19], without however considering the interaction range as a parameter. Further, the most important questions that we shall consider in this paper were not studied there. Several papers [20, 21, 22] investigated the behavior of thermodynamic quantities and local order in Kac spin glasses. Finally a one-dimensional Kac spin glass, with a different (continuous) phase transition was recently studied numerically in [23].

The paper is organized as follows. In Section II we introduce our Kac-XORSAT model, and its variants, and discuss some of their most basic properties in Section III. We investigate thermodynamic quantities (in particular the ground state entropy) in Section IV, and the correlation length divergence in Section V. Finally a discussion of our results is presented in VI, and several technical details are contained in the Appendices.

II. DEFINITION OF THE MODEL

Let us recall that an instance of the k-XORSAT problem is defined by an $M \times L$ matrix binary $\mathbb{H}$, with row weight[2] k and a binary vector $\underline{b}$ of length M. Solving such an instance requires determining whether the linear system

$$\mathbb{H}\underline{x} = \underline{b} \qquad \mathrm{mod}\ \ 2\,, \tag{1}$$

[2] The *row weight* is the number of non-vanishing entries in each row of the matrix.

admits a solution $\underline{x} \in \{0,1\}^L$. This question is equivalent to asking whether the ground state energy of a certain Ising spin model, is zero or not. More precisely, let $\{i_1(a),\dots,i_k(a)\} \subseteq [L]$ denote the indices of the non-zero entries in the a-th row of $\mathbb{H}$, and $J_a = (-1)^{b_a}$ (here $a \in [M]$, and $[n] \equiv \{1,\dots,n\}$). The relevant spin model is defined by letting the energy of configuration $\underline{\sigma} \equiv (\sigma_1,\dots,\sigma_L) \in \{+1,-1\}^L$ be

$$E(\underline{\sigma}) = \sum_{a=1}^{M} \left(1 - J_a \sigma_{i_1(a)} \cdots \sigma_{i_k(a)}\right) . \tag{2}$$

In the following we shall refer to a particular XORSAT instance as to a 'formula' or a 'sample'.

The random k-XORSAT (rXOR) ensemble is defined by letting $\mathbb{H}$ be a uniformly random binary matrix (with dimensions $M \times L$ and row weight k) and $\underline{b}$ a uniformly random vector in $\{0,1\}^L$. It is also useful to consider the *unfrustrated* random k-XORSAT ensemble, defined by setting $\underline{b} = \underline{0}$ (the all 0's vector) deterministically. Such an ensemble exhibits a particularly rich behavior in the 'thermodynamic' limit $L \to \infty$, $M \to \infty$ with $\gamma = M/L$ kept fixed.

The *Kac k-XORSAT* (KacXOR) ensembles add to the above features a one-dimensional (or, in linear algebra terms, a band diagonal) structure. One such ensemble is characterized by the parameters introduced so far, namely k, L, and γ, plus an 'interaction range' R. Unlike in the rXOR ensemble, γ is required to be in $[0,1]$, although generalizations are not difficult. Further, the interaction range is an integer such that $2R+1 \geq k$. Given such parameters, the matrix $\mathbb{H}$ is sampled as follows. Rows of $\mathbb{H}$ are indexed by a subset F of $[L]$: for each $a = 1,\dots,L$, $a \in F$ independently from the others with probability γ. In particular, the number of rows of $\mathbb{H}$, M, is a binomial random variable

$$\mathbb{P}\{M\} = \binom{L}{M} \gamma^M (1-\gamma)^{L-M} . \tag{3}$$

As $L \to \infty$, the number of rows is with high probability[3], close to $L\gamma$. For each $a \in F$, the corresponding row in $\mathbb{H}$ is sampled independently from the others by letting the indices of non-zero entries $(i_1(a),\dots,i_k(a))$ be a uniformly random subset of $\{a-R,\dots,a+R\}$ (i.e. each of the $\binom{2R+1}{k}$ subsets has the same probability). We shall refer to $\{a-R,\dots,a+R\}$ as to the *range* of equation a.

Finally, we let the entries of $\underline{b}$ be indexed by $\mathbb{F}$ as well, and iid random in $\{0,1\}$. As in the case of random XORSAT, some simplification is achieved by considering an unfrustrated ensemble in which $\underline{b} = \underline{0}$.

[3] Following the use in probability theory, we say that something happens *with high probability* (w.h.p.) if its probability approaches 1 in the thermodynamic limit.

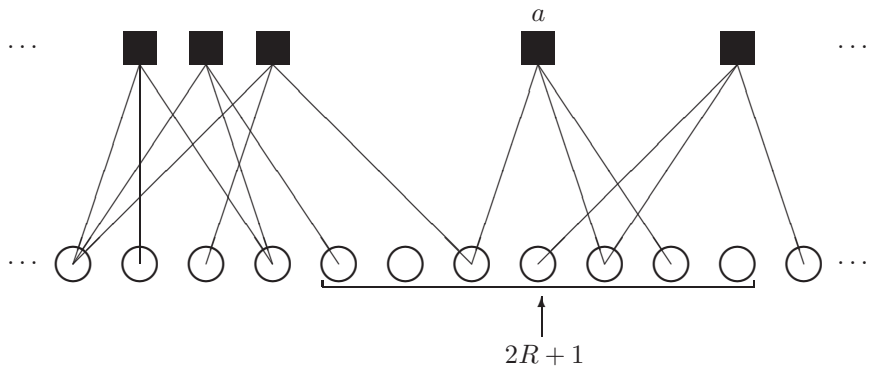

FIG. 1: Factor graph representation of a portion of a KacXOR formula with $k = 3$ and $R = 3$. Empty circles correspond to variables (columns of $\mathbb{H}$) and filled squares to equations (rows of $\mathbb{H}$).

A XORSAT formula admits a natural representation as a factor graph G. This is a bipartite graph including one 'parity check node' for each row in $\mathbb{H}$ (i.e. for each equation in the linear system), and one 'variable node' for each column (i.e. for each variable in the linear system). A parity check and a variable node are connected by an edge if the corresponding entry of $\mathbb{H}$ is non-vanishing. An example of such representation is presented in Fig. 1.

There is still one point of the above definition to be clarified. When $a \le R$ or $a \ge L - R$, the range for equation a is not included in the sets of variable indices, and it might be that $i_l(a) \le 0$ or $i_l(a) > L$. We shall consider two types of boundary conditions. With *periodic boundary conditions*, variable indices are interpreted modulo L. Therefore, if for some index we have $i_l(a) > L$, this is identified with $i_l(a) - L$, while, if $i_l(a) \le 0$, this is identified with $i_l(a) + L$.

In the case of *fixed boundary condition* we will let the set potential indices of row of $\mathbb{H}$ be $\{-R+1, \dots, L+R\}$. Namely F includes each a in this set independently with probability γ. To define a fixed boundary condition, we shall fix a doubly infinite reference configuration[4] $\underline{x}^{(0)} = \{x_i^{(0)} : i \in \mathbb{Z}\}$. If in building row a we get an index $i_l(a) \notin [L]$, the corresponding non-zero entry is not included in $\mathbb{H}$, but rather the value $x^{(0)}_{i_l(a)}$ is added to b_a. This corresponds to fixing $\underline{x} = \underline{x}^{(0)}$ 'outside' $\{1, \dots, L\}$. Finally, we shall agree that, whenever considering the unfrustrated ensemble, the reference configuration will be the all 0's sequence $\underline{x}^{(0)} = \underline{0}$.

[4] Notice that the boundary condition depends on the reference configuration $\underline{x}^{(0)}$ only through $x^{(0)}_{-2R+1}, \dots, x^{(0)}_0$ and $x^{(0)}_{L+1}, \dots, x^{(0)}_{L+2R}$.

III. FRUSTRATED VERSUS UNFRUSTRATED ENSEMBLE

The most important feature of the rXOR ensemble in the large size limit is that it undergoes a SAT-UNSAT phase transition at well defined constraint density $\gamma_{\rm s}(k)$. More precisely, a random XORSAT formula of the type (1) is solvable (SAT) with high probability if $\gamma < \gamma_{\rm s}(k)$, while it is not solvable (UNSAT) for $\gamma > \gamma_{\rm s}(k)$ [13, 14, 15, 16].

It is a convenient feature of XORSAT that this phase transition can be studied by considering uniquely the unfrustrated ensemble. This simplification can be explained through the well-known identity

$$\mathbb{P}\{\,\mathbb{H}\,\underline{x} = \underline{b} \text{ is SAT}\} = \mathbb{E}\left\{2^{L-M}/Z(\mathbb{H})\right\}\,, \tag{4}$$

where $Z(\mathbb{H})$ denotes the number of solutions of the homogeneous linear system $\mathbb{H}\underline{x} = \underline{0} \mod 2$. The identity holds irrespective of distribution of $\mathbb{H}$ provided the right hand side of Eq. (1), i.e. the vector $\underline{b}$, is uniformly random. In order to prove it, it is sufficient to notice that $\mathbb{H}\underline{x} = \underline{b}$ is SAT if and only if $\underline{b}$ is in the image of $\mathbb{H}$. Since the dimension of the image of $\mathbb{H}$ is $\mathrm{rank}(\mathbb{H}) = L - \dim\ker(\mathbb{H})$, this happens with probability $2^{L-\dim\ker(\mathbb{H})}/2^M$. on the other hand, $Z(\mathbb{H}) = 2^{\dim\ker(\mathbb{H})}$. Equation (4) follows by taking expectation with respect to $\mathbb{H}$.

Within the rXOR ensemble, for $\gamma < \gamma_{\rm s}(k)$, $Z(\mathbb{H})$ is tightly concentrated around 2^{L-M}, implying $\mathbb{P}\{\mathbb{H}\,\underline{x} = \underline{b} \text{ is SAT}\} \approx 1$. Viceversa for $\gamma > \gamma_{\rm s}(k)$, typically $Z(\mathbb{H}) \doteq 2^{L\phi(\gamma)}$ (here $\doteq$ denotes equality to leading exponential order), with $\phi(\gamma) > 1-\gamma$ and therefore the formula is SAT with exponentially small probability.

Furthermore, as long as the non-homogeneous solution has at least one solution, its number of solution is independent of $\underline{b}$, and is given by $Z(\mathbb{H})$. Even more, the set of solutions is an affine space obtained by translating the linear space of solutions of the homogeneous system. In other words, conditional to the problem being solvable (which happens with high probability for $\gamma < \gamma_{\rm s}(k)$) the frustrated and unfrustrated ensemble are essentially equivalent.

An important novelty within the KacXOR ensemble is that the linear system (1) is always UNSAT with high probability if we let $L \to \infty$ with γ, R fixed. More precisely, we expect that

$$\mathbb{P}\{\mathbb{H}\,\underline{x} = \underline{b} \text{ is SAT}\} \doteq e^{-L\,\Lambda(\gamma,R)}\,, \tag{5}$$

for some $\Lambda(\gamma, R)$ strictly positive and non-decreasing in γ. This phenomenon was already oserved in [19] for a related model. The basic reason for this behavior is that small subsets of nearby rows of $\mathbb{H}$ have a fair chance of being linearly dependent. If this is the case, the corresponding linear

subsystem is unsolvable with finite probability. In the large L limit, the expected number of such substructures is of order L, and the probability that none is present is exponentially small, thus leading to the above behavior.

It is not difficult to prove the above statement, and indeed to prove lower bounds on the rate $\Lambda(\gamma, R)$ by combinatorial techniques. The basic idea is to select a particular type of substructure that leads to unsatisfiability and estimate the probability that no such substructure is present. The simple such substructure is obtained when two lines of $\mathbb{H}$ coincide but the corresponding entries of $\underline{b}$ do not. Using Janson inequality this yields

$$\Lambda(\gamma, R) \;\geq\; K_1\gamma^2 - K_2\gamma^3(1 - K_0\gamma^2)^{-1}\,, \tag{6}$$

where

$$K_0 \;\equiv\; \binom{2R+1-1}{k}\binom{2R+1}{k}^{-2}, \qquad K_1 \equiv \frac{1}{2\binom{2R+1}{k}^2}\sum_{p=1}^{2R+1-k}\binom{2R+1-p}{k}, \tag{7}$$

$$K_2 \;\equiv\; \frac{3\gamma^3}{8\binom{2R+1}{k}^3}\sum_{p=2}^{2R+1-k}(p-1)\binom{2R+1-p}{k}. \tag{8}$$

Such a lower bound is easily seen to be strictly positive for γ small enough. We refer to Appendix A for a derivation of this formula.

Notice that the lower bound in Eq. (6) vanishes as $1/R^k$ when $R \to \infty$. We expect the same behavior to hold for the actual exponent as long as γ is below the (mean-field) satisfiability threshold $\gamma_{\rm s}(k)$. Explicitely

$$\Lambda(\gamma, R) = \Lambda_1(\gamma)/R^k + O(1/R^{k+1})\,, \tag{9}$$

where $\Lambda_1(\gamma) \uparrow +\infty$ as $\gamma \uparrow \gamma_{\rm s}(k)$.

In the following we shall avoid dealing with the above phenomenon by focusing directly on the unfrustrated ensemble: this will enable us to use efficient linear algebra techniques for numerical computations. There are several justifications for doing this:

1. The two ensembles become equivalent in the Kac limit, which is our main concern here.

2. We are interested in the long distance properties of the model, rather than in the effect of small substructures. We think that the two decouple for large R.

3. Even if the frustrated ensemble is with high probability unsatisfiable, one can always consider 'almost satisfying' configurations. Equivalently, one can study the Boltzmann measure for

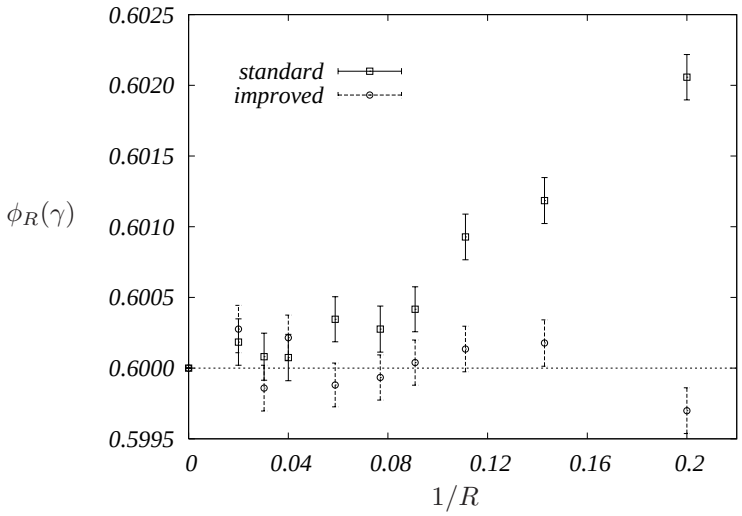

FIG. 2: Ground state entropy density in the thermodynamic limit $\phi_R(\gamma) = \lim_{L\to\infty} \phi_{L,R}(\gamma)$, cf. Eq. (10), for the standard and improved ensembles. Here $k = 3$ and $\gamma = 0.4$. The horizontal line marks the $R \to \infty$ limit $\phi(\gamma) = 0.6$.

the energy (2) at a small non-vanishing temperature T. We expect the effect of small frustrated substructures on the thermodynamics to be small, and indeed vanishing as R^{-k} for large R.

In this perspective, we shall introduce an *improved ensemble* which reduces the effect of small substructured, while keeping the large R behavior unchanged. This is particularly convenient in numerical simulations.

Ideally, one would like to consider a uniform ensemble conditioned on some class of substructures being absent. In practice it can be excedingly difficult to sample matrices $\mathbb{H}$ from such a conditional ensemble. We shall define the improved ensemble by the following sequential procedure. First generate the random set $F \subseteq [L]$ by letting $i \in F$ independently for each $i = 1, \ldots, L$ with probability γ. The set F will index lines of $\mathbb{H}$ as above. Then we choose a uniformly random ordering $(i(1), \ldots, i(M))$ of the elements of F, and generate the corresponding lines of $\mathbb{H}$ following such an order. For each $t = 1, \ldots, M$ we try to generate the line of $\mathbb{H}$ indexed by $i(t)$ by drawing its k non-zero elements uniformly at random in $\{i(t) - R, \ldots, i(t) + R\}$. If the newly generated line has $k - 1$ or k non-vanishing entries in common with one of the previously generated lines $\{i(1), \ldots, i(t-1)\}$, we reject it and re-sample it. We repeat the trial-rejection step for at most 100 times. If no valid line is generate within this round, the whole system generated so-far is rejected and the procedure is re-initiated from scratch.

We shall refer to the first ensemble introduced above as to the *standard*, whenever it will be

necessary to distinguish it from the *improved* ensemble. In Fig. 2 we compare the $R \to \infty$ behavior of the ground state entropy for the two ensembles. Although the limits clearly coincide, the improved ensemble is close to it even for $R = 5$.

IV. GROUND STATE ENTROPY

The simplest thermodynamic quantity that is relevant for the study of the unfrustrated KacXOR problem is the ground state entropy, i.e. the logarithm of the number of solutions of the linear system. Let us denote by $Z(\mathbb{H})$ the number of solutions of the linear system (1) for a random binary matrix $\mathbb{H}$. Then the average entropy density is defined as

$$\phi_{L,R}(\gamma) = \frac{1}{L}\mathbb{E}\log_2 Z(\mathbb{H})\,, \tag{10}$$

In order to compare analytical predictions and numerical data it will be convenient to define the 'subtracted' entropy density $\widehat{\phi}_{L,R}(\gamma) \equiv \phi_{L,R}(\gamma) - \phi_{\rm naive}(\gamma)$, where $\phi_{\rm naive}(\gamma) = 1 - \gamma$. Notice that $\phi_{\rm naive}(\gamma)$ is the naive prediction that would be obtained by assuming the lines of $\mathbb{H}$ to be linearly independent.

Given a matrix $\mathbb{H}$, the corresponding number of solutions takes the form of a partition function

$$Z(\mathbb{H}) = \sum_{\underline{x}} \prod_{a=1}^{L} \psi_a(x_{a-R},\dots,x_{a+R})\,, \tag{11}$$

where (denoting by $\oplus$ the sum modulo 2)

$$\psi_a(x_{a-R},\dots,x_{a+R}) = \begin{cases} \mathbb{I}(x_{i_1(a)} \oplus \dots \oplus x_{i_k(a)} = 0) & \text{if } a \in F, \\ 1 & \text{if } a \notin F. \end{cases} \tag{12}$$

Due to the finite interaction range R, $Z = Z(\mathbb{H})$ can be computed through a transfer matrix algorithm which recursively computes left and right partition functions, respectively $Z_{\to i}$ and $Z_{i\leftarrow}$. These are indexed by $\vec{z} = (z_1,\dots,z_{2R}) \in \{0,1\}^{2R}$, and defined as

$$Z_{\to i}(\vec{z}) \equiv \sum_{\substack{x_1\dots x_i \\ \vec{x}_{i-2R+1}^{\,i} = \vec{z}}} \prod_{a=1}^{i-R} \psi_a(x_{a-R},\dots,a+R)\,, \tag{13}$$

$$Z_{i\leftarrow}(\vec{z}) \equiv \sum_{\substack{x_i\dots x_i \\ \vec{x}_{i}^{\,i+2R-1} = \vec{z}}} \prod_{a=i+R}^{L} \psi_a(x_{a-R},\dots,a+R)\,, \tag{14}$$

where we used the notation $\vec{x}_{j+1}^{\,j+2R} = (x_{j+1},\dots,x_{j+2R})$. A recursion naturally follows

$$Z_{\to(i+1)}(z_1,\dots,z_{2R}) = \sum_{z_0\in\{0,1\}} \psi_{i-R+1}(z_0,z_1,\dots,z_{2R})\, Z_{\to i}(z_0,\dots,z_{2R-1})\,, \tag{15}$$

together for the analogous recursion for $Z_{i\leftarrow}$. It is clear that the total number of solutions can be computed from the constrained partition functions.

The naive transfer matrix algorithm defined by the recursion (15) has complexity that of order $\Theta(L2^{2R})$. This severely limits the interaction ranges that can be treated with this method: in practice we could deal at most with $R = 10 \div 11$, which is far too small to address issues concerning the $R \to \infty$ limit. In order to overcome this problem, we developed a transfer matrix algorithm that, while computing exactly the constrained partition functions, exploits the linear structure of the problem in such a way to reduce the complexity to $\Theta(LR^3)$. Thanks to this approach, we were able to treat systems with $R = 100$ or larger. For details on the algorithm we refer to Appendix B.

We are interested in the double limit $R, L \to \infty$. We shall consider two procedures to define the limit. The first one corresponds to the classical Kac limit, and consists in taking the thermodynamic limit upfront to define

$$\phi_R(\gamma) \equiv \lim_{L\to\infty} \phi_{L,R}(\gamma)\,. \tag{16}$$

Next, we let $R \to \infty$. In Appendix C we will show that $\phi_R(\gamma)$ can be expanded for large R as follows

$$\phi_R(\gamma) = \phi^{(0)}(\gamma) + \frac{1}{2R+1}\,\phi^{(1)}(\gamma) + O\left(\frac{1}{R^2}\right)\,. \tag{17}$$

The leading term gives the mean-field limit and coincides with the ground state entropy density within the rXOR ensemble [15, 16]. It can be expressed in the form $\phi^{(0)}(\gamma) = \max_{\varphi\in[0,1]} \phi^{(0)}(\gamma;\varphi)$, where

$$\phi^{(0)}(\gamma;\varphi) = -\gamma(1-\varphi^k) + k\gamma\varphi^{k-1}(1-\varphi) + e^{-k\gamma\varphi^{k-1}}\,. \tag{18}$$

It is easy to show that the max is achieved for a value of the order parameter φ that satisfies the equation $\varphi = 1 - \exp\{-k\varphi^{k-1}\}$. For $\gamma < \gamma_{\rm s}(k)$, the maximum is at $\varphi = 0$, yielding $\phi^{(0)}(\gamma) = 1-\gamma$. In other words, the the rank of $\mathbb{H}$ is smaller than the maximum possible value by a fraction of order $1/R$. For $\gamma \geq \gamma_{\rm s}(k)$, the maximum is at $\varphi = \varphi_* > 0$ strictly, and $\phi^{(0)}(\gamma) > 1-\gamma$: the rank of $\mathbb{H}$ remains strictly smaller than its maximum possible value even as $R \to \infty$. For instance we have $\gamma_{\rm s}(k) \approx 0.917935$ for $k = 3$.

The first-order contribution $\phi^{(1)}(\gamma)$ is related to fluctuation around the saddle point in an appropriate path integral representation of Eq. (11). Its expression is given in Appendix C.

In Fig. 3 we plot the numerical estimates for the subtracted entropy density $\widehat{\phi}_{L,R}(\gamma)$, as obtained with our transfer matrix algorithm for $R = 25$ and several system sizes. Data points are the result

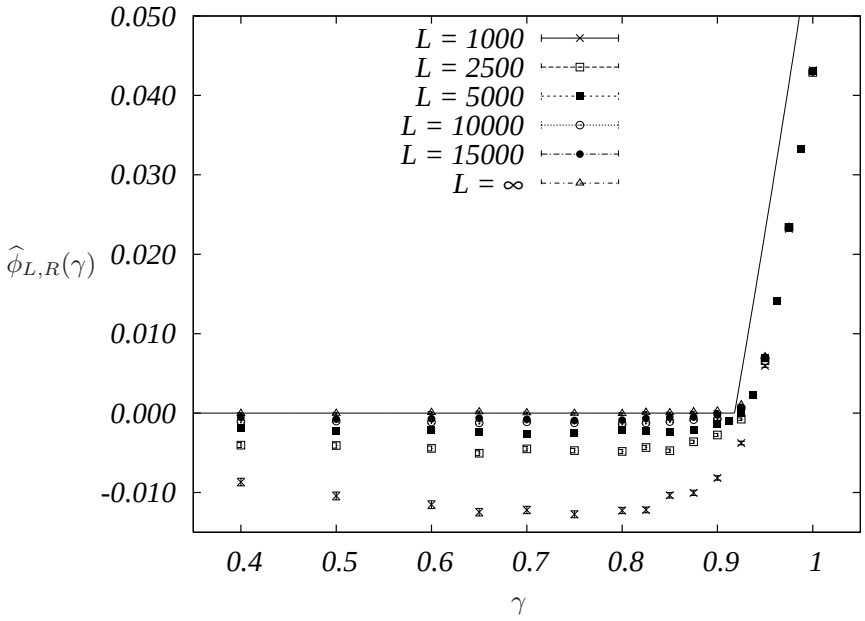

FIG. 3: Subtracted entropy density $\widehat{\phi}_{L,R}(\gamma) = \phi_{L,R}(\gamma) - 1 + \gamma$ for various values of L and $R = 25$ constant. We also plot the result of an $L \to \infty$ extrapolation, and the analytical mean-field prediction $\phi^{(0)}(\gamma) - 1 + \gamma$ (continuous line).

of averaging over 1000 realizations of $\mathbb{H}$ with $k = 3$. The same statistics and value of k will be used in the other numerical experiments below: we shall omit mentioning it again. Further, unless otherwise stated, we will keep using the improved ensemble. We also show the result of an $1/L$ extrapolation to $L = \infty$. The control of the thermodynamic limit is quite good (although corrections at moderate values of L are large). It is clear that the $L = \infty$ extrapolation is not compatible with the mean field prediction, and that $1/R$ corrections must be taken into account.

Figure 4 shows the result of such an $L \to \infty$ extrapolation for several values of R. The data seem to approach the mean field prediction $\phi^{(0)}(\gamma) - 1 + \gamma$ as $R \to \infty$, although the approach is rather slow.

In order to better study the large-R limit, for 4 values of γ we computed the ground state entropy for a wide range of R. The result is compared in Figure 5 with the asymptotic expression (17). In this case we used the standard ensemble which presents larger $1/R$ corrections (computing the first order correction $\phi^{(1)}$ within the improved ensemble is technically much more difficult). It turns out from the analysis in Appendix C that $\phi^{(1)}(\gamma) = 0$ for $\gamma < \gamma_{\rm s}(k)$ while $\phi^{(1)}(\gamma) \neq 0$ for $\gamma \geq \gamma_{\rm s}(k)$. Our data confirm this behavior. Further, although $O(R^{-2})$ contributions are rather large, the leading $1/R$ correction to mean field does indeed match the analytical prediction.

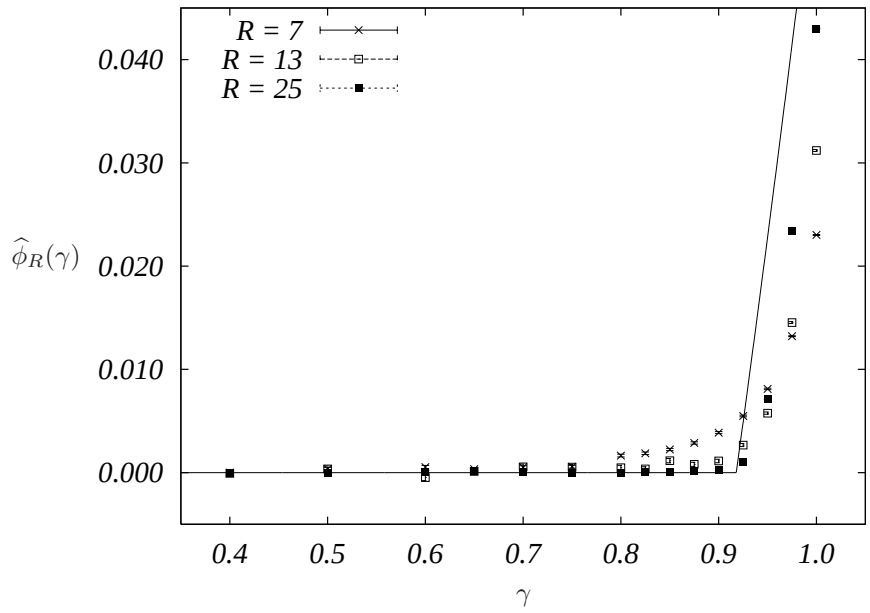

FIG. 4: Subtracted entropy density in the thermodynamic limit: $\widehat{\phi}_R(\gamma) = \phi_R(\gamma) - 1 + \gamma$ for various values of R, together with the mean field prediction $\phi^{(0)}(\gamma) - 1 + \gamma$ (continuous line).

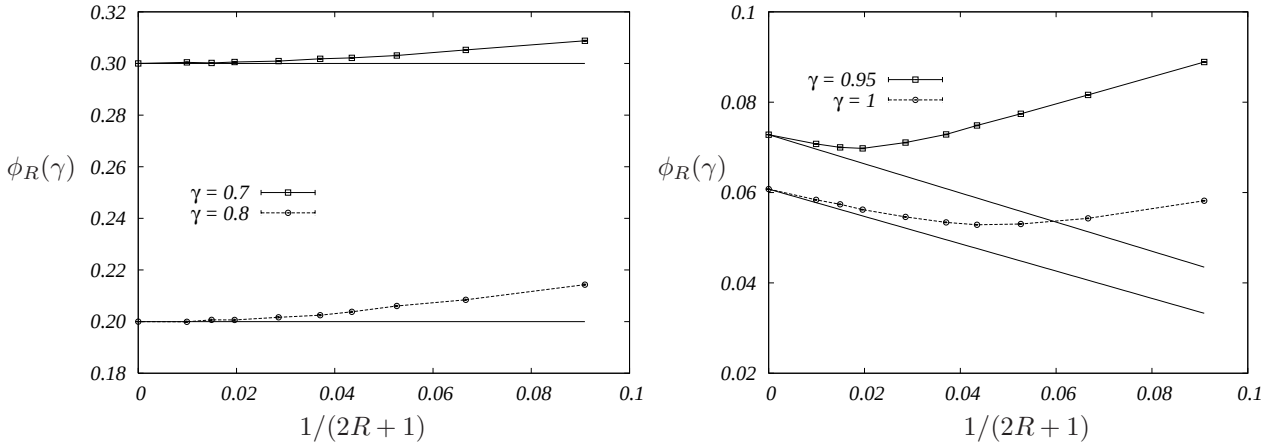

FIG. 5: Ground state entropy density $\phi_R(\gamma)$, extrapolated to the thermodynamic limit, versus the inverse interaction range $1/(2R+1)$ for various values of γ. Straight lines correspond to the analytic prediction, cf. Eq. (17).

The second limit we shall consider is $L, R \to \infty$ with $\ell \equiv L/R$ kept fixed. We thus define

$$\phi^*_\ell(\gamma) \equiv \lim_{R\to\infty} \phi_{R\ell,R}(\gamma)\,. \tag{19}$$

This is the mean-field limit for a system of finite size. The limit can be computed exactly by maximizing an appropriate action functional over a position-dependent order parameter. More

precisely we have $\phi^*_\ell(\gamma) = \max_\varphi A[\varphi]$, where $\varphi : [0,\ell] \to \mathbb{R}$ is the order parameter, and

$$A[\varphi] = \frac{1}{\ell}\int_0^\ell \left\{\gamma - k\gamma\varphi(z)^{k-1} + (k-1)\gamma\varphi(z)^k - \exp\left[-\frac{k\gamma}{2}\int_{-1}^{1}\varphi(z+u)^{k-1}\mathrm{d}u\right]\right\}\mathrm{d}z\,. \qquad (20)$$

By differentiating with respect to φ, we obtain the mean-field equation

$$\varphi(z) = 1 - \frac{1}{2}\int_{-1}^{+1}\exp\left\{-\frac{k\gamma}{2}\int_{-1}^{+1}\varphi(z+u+v)^{k-1}\mathrm{d}v\right\}\mathrm{d}u\,. \qquad (21)$$

We refer to Appendix C for a derivation of these formulae and limit ourselves to discuss their consequences here.

In the case of a homogeneous order parameter $\varphi(z) = \varphi$ independent of z, Eq. (21) is satisfied if φ if a solution of the standard mean field equation, $\varphi = 1 - \exp\{-k\gamma\varphi^{k-1}\}$. The action (20) then reduces to the mean field free energy Eq. (18).

In the general case the order parameter $\varphi(z)$ has a simple interpretation. Consider the linear system $\mathbb{H}\underline{x} = \underline{0} \mod 2$, and let $i \in \{1,\dots,L\}$. Then, one of the following must happen: either $x_i = 0$ in all of the solutions; or $x_i = 0$ in half of the solutions and $x_i = 1$ in the other half. We shall call x_i (or, sometimes, i) a frozen variable in the first case, and a free variable in the second one. Given $z \in [0,\ell]$, the number of frozen variables with $i \in [Rz, R(z+\mathrm{d}z)]$ in a typical random linear system from our ensemble, is about $R\varphi(z)\,\mathrm{d}z$. Equivalently, the probability for x_i, $i = \lfloor Rz \rfloor$, to be frozen converges to $\varphi(z)$.

We shall come back to this interpretation in the next Section, while using it here to derive the appropriate boundary conditions for Eq. (21). If the linear system is defined with periodic boundary conditions, then we have to use periodic boundary conditions in Eq. (21) as well, namely $\varphi(z+\ell) = \varphi(z)$. If on the other hand we adopt fixed boundary conditions with respect to the reference solution $\underline{x}^{(0)} = \underline{0}$, then we have to impose $\varphi(z) = 1$ for $z \le 0$ and $z \ge \ell$ in Eq. (21). As a consequence, the homogeneous solution is no longer relevant in this case.

Once the boundary conditions have been estabilished, Eq. (21) can be solved numerically by iteration (after discretizing z on a sufficiently fine mesh). In the regime in which multiple fixed points exist, the relevant one is obtained by maximizing the action.

The result of such a computation is compared in Fig. 6 with the outcome of numerical simulations. The agreement is good already at moderate interaction ranges. The main effect of a finite size is a decrease in the number of solutions due to the fact that variables close to the boundary are more highly constrained (and thus more likely to be frozen). This effect is accurately reproduced by the analytical calculation.

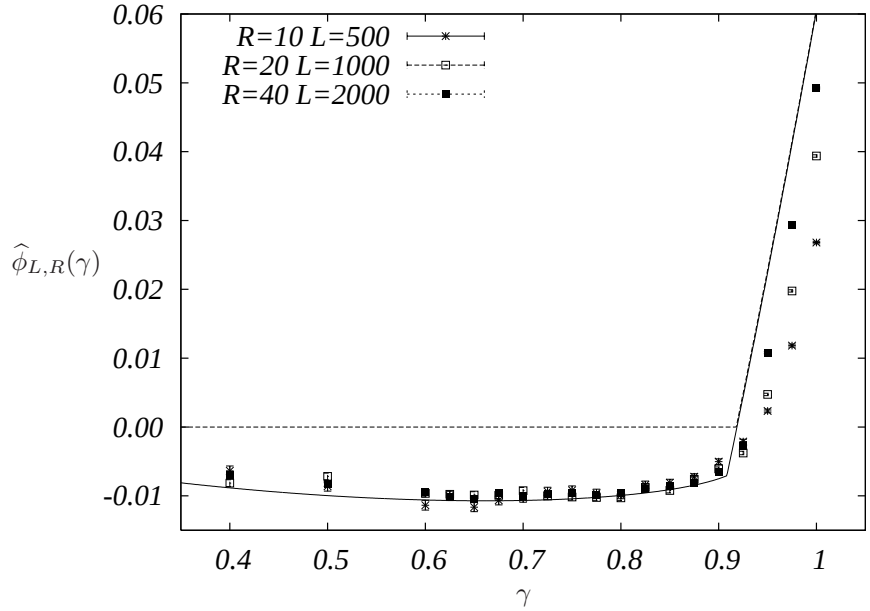

FIG. 6: Subtracted entropy density $\widehat{\phi}_{L,R}(\gamma) = \phi_{L,R}(\gamma) - 1 + \gamma$ as a function of γ for several at $\ell = L/R = 50$ fixed. The continuous line corresponds to the analytical prediction $\phi^*_\ell(\gamma) - 1 + \gamma$ in the $R \to \infty$ limit.

V. POINT-TO-SET CORRELATION FUNCTION

As we have seen in the previous Section, the thermodynamic behavior of the KacXOR ensemble at finite R carries several traces of the mean field limit. Here we want to investigate some structural features of the uniform measure over solutions of the linear system:

$$\mu(\underline{x}) = \frac{1}{Z}\,\mathbb{I}(\mathbb{H}\,\underline{x} = \underline{0}) \equiv \frac{1}{Z}\,\prod_a \psi_a(x_{a-R}, \dots, x_{a+R})\,. \tag{22}$$

In particular, we want to understand whether the mean field ergodicity breaking transition shows up in the long range correlations of this measure, as predicted within the mosaic state scenario.

It is expected that the long range order emerging at a glass transition cannot be probed through ordinary point-to-point correlations functions, and that point-to-set correlation functions have to be used instead [11]. These can be defined through the following "experiment" [10] (we refer here to the one-dimensional case we are studying). Consider a large sample $L \gg R$, let i be a node in its bulk: $i \gg R$, $L - i \gg R$, and $\underline{x}^*$ a 'reference' configuration sampled from the measure $\mu(\,\cdot\,)$. Then fix some $1 \le \widetilde{L} \ll L$, and consider a second configuration that is forced to coincide with $\underline{x}^*$ on sites j with $|j - i| > \widetilde{L}$, and free otherwise, and compute the probability that $x_i \neq x^*_i$. The expectation of this probability with respect to $\underline{x}^*$ and the sample realization yields the desired point-to-set correlation. In formulae, if we let $\mathsf{B}(i, \widetilde{L}) = \{j : |i - j| \le \widetilde{L}\}$ be the box of size $2\widetilde{L} + 1$

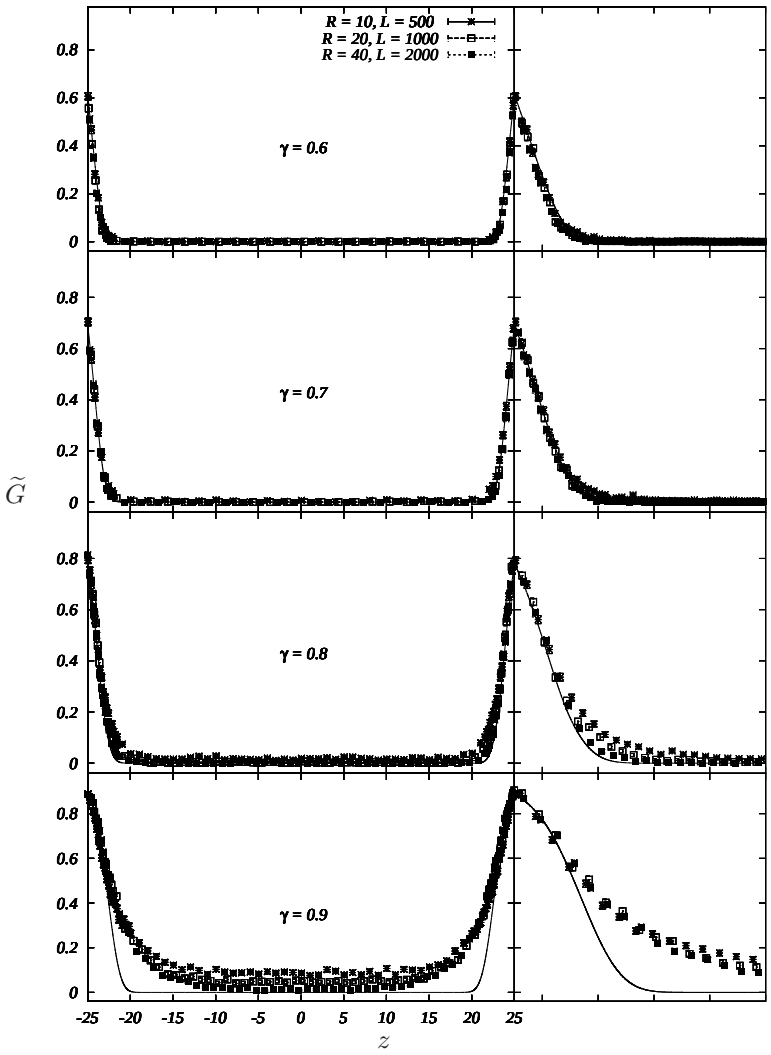

FIG. 7: Correlation $\widetilde{G}(n;\widetilde{L},R,\gamma)$ between the boundary of a box of size $2\widetilde{L}+1=2\ell R+1$ and a point in its interior (at distance $n=zR$ from the center). In the right frames: blow-up of the region near the boundary. The continuous line (partially hidden by data points) corresponds to the analytic prediction obtained by solving Eq. (21).

around i, and $\overline{\mathsf{B}}(i,\widetilde{L})$ its complement, we define

$$\widetilde{G}(\widetilde{L},R,\gamma)\equiv\mathbb{E}\{1-2\mu_{i|\overline{\mathsf{B}}(i,\widetilde{L})}(x_i\neq x_i^*|\underline{x}^*_{\overline{\mathsf{B}}(i,\widetilde{L})})\}\,. \tag{23}$$

Here the thermodynamic limit $L\to\infty$ is assumed to be taken at the outset, $\mathbb{E}$ denotes expectation both with respect to the matrix $\mathbb{H}$ and the reference configuration $\underline{x}^*$, and the redefinition $1-2(\cdots)$ is for future convenience.

The linear structure of our problem implies two simplifications. First, the conditional probability appearing in Eq. (23) is indeed independent of $\underline{x}^*$ (that can be 'gauged away'). Therefore we can fix $\underline{x}^*=\underline{0}$ and eliminate the expectation over the reference configuration $\underline{x}^*$. The resulting conditional

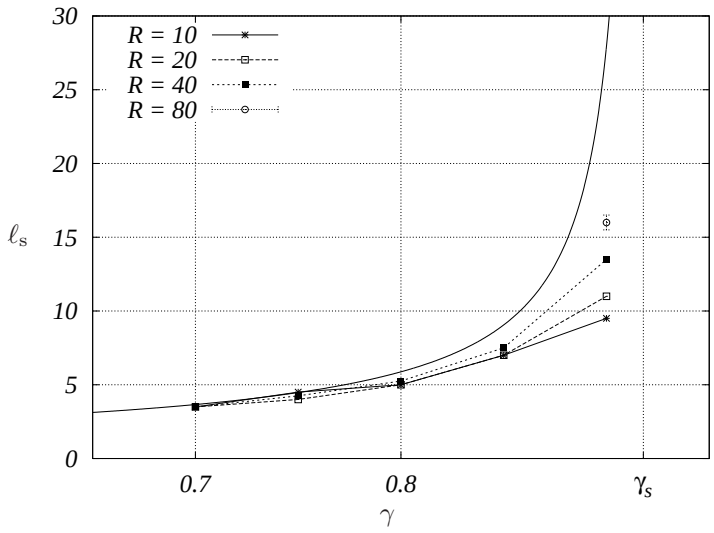

FIG. 8: Point-to-set correlation length in units of the interaction range R. The continuous line corresponds to the analytic prediction for $R \to \infty$ and diverges at the glass transition $\gamma_s(k=3) \approx 0.917935$.

measure is just the distribution of a system with fixed boundary conditions $\underline{0}$ as discussed in the previous Section. This implies a second simplification (already noticed above). The conditional probability $\mu_{i|\overline{\mathsf{B}}(i\widetilde{L})}(x_i \neq 0 | \underline{x}^*_{\overline{\mathsf{B}}(i\widetilde{L})} = \underline{0})$ can take value 1/2 (if x_i is 'free') or 0 (if it is 'frozen'). We thus get

$$\widetilde{G}(\widetilde{L}, R, \gamma) = \mathbb{P}_{\widetilde{L}}\{\, x_{\widetilde{L}+1} \text{ is free}\,\}\,.$$

Here $\mathbb{P}_{\widetilde{L}}$ denotes probability with respect to a matrix $\mathbb{H}$ with $2\widetilde{L}+1$ columns and fixed 0 boundary conditions. In fact it is interesting to generalize the above definition and consider the correlation between any point inside the box of size $2\widetilde{L}$ and its boundary

$$\widetilde{G}(n; \widetilde{L}, R, \gamma) \equiv \mathbb{P}_{\widetilde{L}}\{\, x_{\widetilde{L}+1+n} \text{ is free}\,\}\,.$$

The original definition is recovered for $n = 0$.

We expect $\widetilde{G}(n; \widetilde{L}, R, \gamma)$ to be close to 1 when n approaches the boundaries of the box (i.e. $n \approx \widetilde{L}$ or $n \approx -\widetilde{L}$) and to decrease in the interior. If the box is large enough, it will approach its thermodynamic value, independent of the boundary condition, near the center (for $n \approx 0$). In Figure 7 we show the outcomes of a numerical calculation of $\widetilde{G}(n; \widetilde{L}, R, \gamma)$ for several values of its parameters.

We are particularly interested in the mean field limit. This is obtained by defining

$$G(z; \ell, \gamma) \equiv \lim_{R\to\infty} \widetilde{G}(n = Rz; \widetilde{L} = R\ell, R, \gamma)\,, \tag{24}$$

that is by measuring lengths in terms of the interaction range and letting $R \to \infty$. In agrement with the interpretation of the previous section, we expect $G(z;\ell,\gamma) = \varphi(z)$, where $\varphi(z)$ solves Eq. (21) with boundary condition $\varphi(z) = 1$ for $z \leq -\ell$, and for $z \geq \ell$. The comparison with numerical data in Fig. 7 is satisfactory although the convergence to the $R \to \infty$ limit gets slower and slower as $\gamma_{\rm s}(3) \approx 0.917935$ is approached.

The point-to-set correlation function can be used to define a correlation length, namely the smallest box size such that the correlation is below a pre-estabilished constant ε. Here we will choose[5] $\varepsilon = 1/2$. In formulae

$$\ell_{\rm s}(\gamma, R) = \min\{\ \ell\ :\ \widetilde{G}(\widetilde{L} = R\ell, R, \gamma) \leq 1/2\ \}\,. \tag{25}$$

An analytical prediction in the $R \to \infty$ limit can be obtained by solving Eq. (21) with boundary conditions $\varphi(z) = 1$ for $z \notin [-\ell, \ell]$. The resulting length can be shown to diverge at $\gamma_{\rm s}$ as $\ell_{\rm s}(\gamma, R = \infty) \sim (\gamma_{\rm s} - \gamma)^{-1}$, in agreement with the mosaic picture (indeed $\Sigma(\gamma) \sim (\gamma_{\rm s} - \gamma)$ close to the transition).

In Fig. 8 we compare this prediction with the estimates from numerical simulations at finite R. These two are clearly consistent, although the convergence is rather slow in the critical regime.

VI. DISCUSSION

We defined a simple ensemble of constraint satisfaction problems (more precisely, an ensemble of linear problems over integers modulo 2), with one-dimensional Kac structure. The model is exactly soluble for infinite interaction range $R \to \infty$ and exhibits in this limit a glassy phase with an exponential number of pure states and a SAT-UNSAT transition.

Mean field theory (as interpreted within the mosaic picture) seems to describe the behavior of the system at moderately large R. Indeed we were able to get quantitative predictions by taking into account the principal modifications of naive mean-field theory, namely a position-dependent order parameter, and $1/R$ corrections. In particular we checked for the first time the divergence of the mosaic length scale in a concrete model, by comparing the the result of a controlled approximation (large R limit) with exact numerical calculations.

We think the KacXOR model can be a useful playground for many ideas developed in the physics of glasses. Among several interesting research directions, one might consider: (i) Studying

[5] Any strictly positive constant below the Edwards-Anderson parameter (in this case given by the largest solution of $\varphi = 1 - \exp\{-k\gamma\varphi^{k-1}\}$) should provide an equivalent definition.

the frustrated ensemble (corresponding to an inhomogeneous linear system); (*ii*) Introducing a non-vanishing temperature and studying the corresponding Boltzmann distribution; (*iii*) Studying the behavior of Glauber dynamics, and in particular the relation between relaxation time and mosaic length scale.

On a different theme, ensembles of random constraint satisfaction problems have been recurrently used to test heuristic algorithms [13]. Such tests have limited scope because in practical applications instances are often structured. It might be insightful therefore to consider ensembles with some tunable 'structure parameter', such as the interaction range R in the present model.

APPENDIX A: COUNTING SMALL SUBSTRUCTURES

Consider the random linear system $\mathbb{H}\underline{x} = \underline{b}$ defined in Section II. If two lines $i, j \in F$ in $\mathbb{H}$ are equal, while the corresponding entries in $\underline{b}$ (namely b_i and b_j) are different, then the system has no solution. We call such a pair (i, j) a 'bad pair,' and will write $B_{ij} = 1$ if (i, j) is bad, and $B_{ij} = 0$ otherwise. Therefore

$$\mathbb{P}\left\{\mathbb{H}\underline{x} = \underline{b} \text{ is SAT}\right\} \le \mathbb{P}\{\cap_{(i,j)}[B_{ij} = 0]\}\,, \tag{A1}$$

where the intersection ranges over i, j such that $i < j \le i + 2R + 1 - k$. Let $B = \sum_{(ij)} B_{ij}$ be the number of bad pairs. In order to bound the right hand side above, we use Janson's inequality [24], which implies

$$\mathbb{P}\left\{\mathbb{H}\underline{x} = \underline{b} \text{ is SAT}\right\} \le \exp\left\{-\mathbb{E}[B] + \Delta/2(1-\epsilon)\right\}\,. \tag{A2}$$

Here

$$\epsilon = \sup_{(ij)} \mathbb{E}[B_{ij}]\,, \qquad \Delta = \sum_{(ij)\sim(lm)} \mathbb{E}[B_{ij}B_{lm}]\,. \tag{A3}$$

where the sum over $(ij) \sim (lm)$ runs over all the couples of distinct pairs (ij) and (lm) such that B_{ij} and B_{lm} are not independent.

It is easy to realize that both $\mathbb{E}[B]$ and Δ are of order $\Theta(L)$ since they are sums of $\Theta(L)$ positive terms. Since we are only interested in the coefficient of the order L term, we shall always consider pairs (ij) in the bulk. Then we have

$$\mathbb{E}[B_{ij}] = \frac{\gamma^2}{2\binom{2R+1}{k}^2}\binom{2R+1-|i-j|}{k}\,. \tag{A4}$$

The factor γ^2 has to be included for having $i,j \in F$ (the two equations must present), $\binom{2R+1-|i-j|}{k}/\binom{2R+1}{k}^2$ is the probability that the two lines in $\mathbb{H}$ coincide, and $1/2$ is the probability that $b_i \neq b_j$

Since the above expression is maximized for $|i-j|=1$, we have $\epsilon = K_0\gamma^2$, with K_0 as in Eq. (7). Further, by summing over i,j we obtain $\mathbb{E}[B] = K_1\gamma^2 L + O(1)$ for $L \to \infty$.

As for the term Δ, the only non-vanishing contribution comes form the case in which there are three distinct indices among $\{i,j,l,m\}$. If we denote by $\underline{h}_n$ the line inedexed by n in $\mathbb{H}$, we get

$$\Delta = \frac{3}{4}\sum_{i<j<l} \mathbb{P}\{i,j,l \in F \text{ and } \underline{h}_i = \underline{h}_j = \underline{h}_l\}\,. \tag{A5}$$

The factor 3 counts the number of different couples of pairs in $\{i,j,l\}$ and $1/4$ is the probability that the corresponding entries in $\underline{b}$ are different. By computing the above probability and summing over i,j,l we get $\Delta = K_2\gamma^3 L + O(1)$, thus proving Eq. (6).

APPENDIX B: POLYNOMIAL TRANSFER MATRIX ALGORITHM

Consider the constrained partition function (13) and the corresponding transfer matrix recursion (15). In this Appendix we shall consider only left-to-right iterations and drop the arrow $\rightarrow$ in subscripts. We shall further set $n = 2R$ and use the vector notation $\vec{x}_{j+1}^{\,j+n} = (x_{j+1},\ldots,x_{j+n})$.

The constrained partition function $Z_i(\vec{z})$ is just the number of solutions in an inhomogeneous linear system, obtained by retaining the lines of $\mathbb{H}$ with index in $\{1,\ldots,i-R\}$ (and the corresponding equations), and adding the n equations $x_{i-n+1} = z_1$, $\ldots$, $x_i = z_n$. As a consequence, for all the choices of $\vec{z}$ such that this linear system has a solution, it has the same number of solutions as corresponding homogeneous system. Further, the number of solutions of the homogeneous system is a power of 2 (because it is the size of a linear space over $\mathbb{Z}_2$). Finally, the vectors $\vec{z}$ for which a solution exists form a linear space. Therefore, there exists a binary matrix $\mathbf{A}_i$ and an integer Φ_i such that

$$Z_i(\vec{z}) = \begin{cases} 2^{\Phi_i} & \text{if } \mathbf{A}_i\vec{z} = \vec{0}, \\ 0 & \text{otherwise.} \end{cases} \tag{B1}$$

The matrix $\mathbf{A}_i$ can always be chosen as an $n\times n$ matrix by eventually eliminating linearly dependent lines.

We therefore reduced the memory requirements from $\Theta(2^n)$ to $\Theta(n^2)$. We have now to show that the $\mathbf{A}_i$ and Φ_i can be computed recursively in polynomial time as well. Consider the recursion (15) and let $a_i = (a_{i,1},\ldots,a_{i,n+1})$ be the binary vector defined as follows. If $i-R+1 \notin F$ (the

new equation added in the recursion is not present), then $a_i \equiv \mathtt{0}$. Otherwise, $a_{i,j} \equiv H_{i-R+1,i-2R+j}$ (a_i encodes the newly added line of $\mathbb{H}$, properly shifted). Then define the $(n+1)\times(n+1)$ matrix $\mathbf{B}_i$ as follows

$$\mathbf{B}_i = \begin{pmatrix} & & 0 \\ & \mathbf{A}_i & \vdots \\ & & 0 \\ a_{i,1} & \cdots\cdots & a_{i,n+1} \end{pmatrix}. \tag{B2}$$

Denote by $\mathbf{b}_i$ the first column of $\mathbf{B}_i$ (i.e. a column vector) and by $\widetilde{\mathbf{B}}_i$ the $(n+1)\times n$ matrix formed by its last n columns. By using Eq. (B1) the recursion (15) can be written as

$$Z_{i+1}(\vec{z}) = \; 2^{\Phi_i} \sum_{z_0\in\{\mathtt{0},\mathtt{1}\}} \mathbb{I}\big(\mathbf{b}_i\, z_0 + \widetilde{\mathbf{B}}_i\, \vec{z} = \mathtt{0}\big)\,. \tag{B3}$$

Let us now consider two cases:

- If $\mathbf{b}_i = \mathtt{0}$, then we get immediately the form (B1) for Z_{i+1}, by letting $\Phi_{i+1} = \Phi_i + 1$, and $\mathbf{A}_{i+1}$ the matrix obtained by eliminating linear dependencies among rows of $\widetilde{\mathbf{B}}_i$.

- If $\mathbf{b}_i \neq \mathtt{0}$, then there exists at least one vector $\widehat{\mathbf{b}}_i$ of dimension $(n+1)$ such that $\widehat{\mathbf{b}}_i^T\mathbf{b}_i = \mathtt{1}$ mod 2. The only non-vanishing term in the sum (B3) is therefore obtained for $z_0 = -\widehat{\mathbf{b}}_i^T\widetilde{\mathbf{B}}_i\vec{z}$ mod 2. Substituting this value of z_0, we obtain that Z_{i+1} can again be written in the form (B1). The new matrix $\mathbf{A}_{i+1}$ is obtained by eliminating linearly dependent rows from $(\mathbf{1} - \mathbf{b}_i\widehat{\mathbf{b}}_i^T)\widetilde{\mathbf{B}}_i$, while the number of solutions is updated by $\Phi_{i+1} = \Phi_i$.

In practice we found more convenient to reduce $\mathbf{B}_i$ in upper triangular form by gaussian elimination before computing $\mathbf{A}_{i+1}$ and Φ_{i+1} as just described.

The initialization of the above recursion depends on the choice of boundary conditions. When using fixed boundary conditions with reference solution $\underline{x}^{(0)} = \underline{\mathtt{0}}$, we set $\mathbf{A}_0 = \mathbf{1}$ and $\Phi_0 = 0$.

It is clear that the above procedure can be implemented in a time that is polynomial in the interaction range. Indeed the most complex operation to be performed, consists in eliminating linearly dependent lines from the matrix $\widetilde{\mathbf{B}}_i$, or $(\mathbf{1} - \mathbf{b}_i\widehat{\mathbf{b}}_i^T)\widetilde{\mathbf{B}}_i$. This can be done via gaussian elimination in time $O(R^3)$. The total complexity is therefore $O(LR^3)$.

APPENDIX C: ANALYTICAL CALCULATIONS

1. Replicas

In order to compute the ground state entropy and the point-to-set correlation function, we shall follow the replica approach, see [25]. Each site $i \in \{1, \dots, L\}$ thus carries n binary variables $\vec{x}_i = (x_i^1, \dots, x_i^n)$ corresponding to the n replicas.

Let us consider first a particularly simple instance consisting of a single equation labeled by $i \in F$ and $2R+1$ variables on sites $j \in \{i-R, \dots, i+R\}$. Denote by $\overline{c}_i(\vec{x})$ the fraction of nodes j such that $\vec{x}_j = \vec{x}$. In formulae

$$\overline{c}_i(\vec{x}) = \frac{1}{2R+1} \sum_{j=i-R}^{i+R} \mathbb{I}(\vec{x}_j = \vec{x}) \,. \tag{C1}$$

The probability that a randomly sampled equation at i (with range $\{1-R, \dots, i+R\}$) is satisfied by all of the n replicas, is a function of $\overline{c}_i$, call it $\mathbb{F}_{k,R}(\overline{c}_i)$. For large R it is easy to show that

$$\mathbb{F}_{k,R}(\overline{c}) = \mathbb{J}_k(\overline{c}) + \frac{1}{2R+1} \binom{k}{2} [\mathbb{J}_k(\overline{c}) - \mathbb{J}_{k-2}(\overline{c})] + O(R^{-2}) \,, \tag{C2}$$

where

$$\mathbb{J}_l(\overline{c}) \equiv \sum_{\vec{x}_1 \dots \vec{x}_l} \prod_{a=1}^{n} \mathbb{I}(x_1^a \oplus \cdots \oplus x_l^a = \mathtt{0})\, \overline{c}(\vec{x}_1) \cdots \overline{c}(\vec{x}_l) \,. \tag{C3}$$

Consider now the full linear system and the partition function (11). We shall implicitly assume periodic boundary conditions in order to lighten the notations. Fixed boundary conditions can be recovered by properly constraining the expressions that we will derive. It follows from the above that

$$\mathbb{E}\{Z^n\} = \sum_{\{\vec{x}_i\}} \prod_{i=1}^{L} [1 - \gamma + \gamma \mathbb{F}_{k,R}(\overline{c}_i)] \,. \tag{C4}$$

Next we introduce two variables $\lambda_i(\vec{x})$, $c_i(\vec{x})$ indexed by $\vec{x} \in \{\mathtt{0}, \mathtt{1}\}^n$ for each $i \in \{1, \dots, L\}$, using the identity

$$1 = \int \mathrm{d}c_i(\vec{x}) \int_{-i\infty}^{+i\infty} \frac{\mathrm{d}\lambda_i(\vec{x})}{2\pi i} \exp\{-\lambda_i(\vec{x})(c_i(\vec{x}) - \overline{c}_i(\vec{x}))\} \,. \tag{C5}$$

This allows to perform the sum over $\vec{x}_i$ in Eq. (C4). If we expand the resulting expression for large R we get, after some lengthy but straightforward calculations,

$$\mathbb{E}\{Z^n\} = \int \mathrm{d}c_i(\vec{x}) \int_{-i\infty}^{+i\infty} \frac{\mathrm{d}\lambda_i(\vec{x})}{2\pi i} \exp\Big\{ -(2R+1)\, S_0[c, \lambda] - S_1[\, c\,] + O(1/R) \Big\} \,, \tag{C6}$$

where

$$S_0[c,\lambda] \;=\; \frac{1}{2R+1}\sum_{i=1}^{L}\left\{-\log[1-\gamma+\gamma\mathbb{J}_k(c_i)]+\sum_{\vec{x}}\lambda_i(\vec{x})c_i(\vec{x})-\log\left[\sum_{\vec{x}} e^{\sum_{j\in\mathsf{D}(i)}\frac{\lambda_j(\vec{x})}{2R+1}}\right]\right\}, \quad \text{(C7)}$$

$$S_1[\,c\,] \;=\; \frac{1}{2R+1}\sum_{i=1}^{L}\gamma\binom{k}{2}\frac{\mathbb{J}_{k-2}(c_i)-\mathbb{J}_k(c_i)}{1-\gamma+\gamma\mathbb{J}_k(c_i)}\,, \quad \text{(C8)}$$

and we introduced the notation $\mathsf{D}(i)\equiv\{j:\,|i-j|\le R\}$.

2. Mean field limit

In the $R\to\infty$ limit, the integral (C6) is dominated by the saddle points of $S_0[c,\lambda]$. We neglect for the moment the correction given by $S_1[c]$, and look for a saddle point of the type

$$c_i(\vec{x})=\varphi_i\,\delta_{\vec{x},\vec{x}_0}+\frac{1}{2^n}(1-\varphi_i)\,,\qquad \lambda_i(\vec{x})=\omega_i\,\delta_{\vec{x},\vec{x}_0}+\frac{1}{2^n}\omega_i^0\,, \quad \text{(C9)}$$

where $\vec{x}_0\equiv(\mathsf{0},\mathsf{0},\dots,\mathsf{0})$ and $\delta_{\vec{x},\vec{y}}$ is the n-dimensional Kronecker delta function. There are several reasons for this Ansatz: (*i*) The algebra of functions of the form $f(\underline{x})=f_0+f_1\delta_{\vec{x},\vec{x}_0}$ is closed; (*ii*) This ansatz is known to give the correct thermodynamic behavior for the rXOR ensemble (i.e. in the mean-field limit); (*iii*) Although it is replica symmetric, it yields the correct one-step replica symmetry breaking physics (it is a peculiarity of XORSAT that replica symmetry can be explicitely broken).

By substituting in Eq. (C7) and letting $n\to 0$, we get $S_0[c,\lambda]=A_0[\varphi,\omega]n\log 2+O(n^2)$ where

$$A_0[\varphi,\omega]=\frac{1}{2R+1}\sum_{i=1}^{L}\left\{\gamma(1-\varphi_i^k)-\omega_i(1-\varphi_i)-e^{-\sum_{j\in\mathsf{D}(i)}\frac{\omega_j}{2R+1}}\right\}. \quad \text{(C10)}$$

By differentiating with respect to φ_i and ω_i we get the saddle point equations

$$\varphi_i=1-\frac{1}{2R+1}\sum_{j\in\mathsf{D}(i)}e^{-\sum_{l\in\mathsf{D}(j)}\frac{\omega_l}{2R+1}}\,,\qquad \omega_i=k\gamma\varphi_i^{k-1}\,. \quad \text{(C11)}$$

The second of these equations can be used to eliminate ω_i from the action.

If we finally assume that φ_i only depends on i on a scale of order R, we can set (with an abuse of notation) $\varphi_i=\varphi(i/R)$ and let $R\to\infty$ with $L=\ell R$, thus getting Eqs. (19) to (21).

3. $1/R$ corrections

In computing the $1/R$ corrections we shall assume the system to be homogeneous. For instance we can think of imposing periodic boundary conditions, or letting $L\to\infty$ at the outset. As a

consequence, in the leading order calculation we have $\varphi_i = \varphi$ independent of i and thus $A_0[\varphi, \omega] = L\phi^{(0)}(\gamma; \varphi)$ with $\phi^{(0)}(\gamma; \varphi)$ as in Eq. (18). Hereafter φ will denote a solution of the mean field equation $\varphi = 1 - \exp\{-k\gamma\varphi^{k-1}\}$, and we let $\omega = k\gamma\varphi^{k-1}$.

There are two contribution to order $1/R$. The first one comes from the correction to the action and is easy to compute. Substituting our Ansatz in Eq. (C8) and proceeding as in the previous Section we get $S_1[c] = A_1[\varphi]\, n \log 2 + O(n^2)$, where

$$A_1[\varphi] = \frac{L}{2R+1}\gamma\binom{k}{2}\varphi^{k-1}(1-\varphi^2)\,. \tag{C12}$$

The second term comes from gaussian fluctuations around the saddle point. Let $c_i^*(\vec{x})$, $\lambda_i^*(\vec{x})$ denote the saddle point (C9) and define

$$c_i(\vec{x}) = c_i^*(\vec{x}) + \nu_i(\vec{x})\,, \quad \lambda_i(\vec{x}) = \lambda_i^*(\vec{x}) + \xi_i(\vec{x})\,. \tag{C13}$$

By expanding $S_0[c, \lambda]$ to second order around its saddle point, we get

$$\begin{aligned} S_0[c,\lambda] \;=\; & S_0[c^*,\lambda^*] + \frac{1}{2(2R+1)}\sum_{i,j}\sum_{\vec{x},\vec{y}} \qquad\qquad (C14)\\ & \Big\{A(\vec{x},\vec{y})\delta_{ij}\nu_i(\vec{x})\nu_j(\vec{y}) + \delta_{\vec{x},\vec{y}}\delta_{i,j}[\nu_i(\vec{x})\xi_i(\vec{y}) + \xi_i(\vec{x})\nu_j(\vec{y})] - B(\vec{x},\vec{y})\kappa_R(i-j)\xi_i(\vec{x})\xi_j(\vec{y})\Big\}\,, \end{aligned}$$

where

$$\kappa_R(i) = \begin{cases} (2R+1-|i|)/(2R+1)^2 & \text{for } |i| \le 2R+1, \\ 0 & \text{otherwise.} \end{cases} \tag{C15}$$

The coefficients appearing in Eq. (C14) have the form

$$A(\vec{x},\vec{y}) \;=\; A_1 + A_2\delta_{\vec{x},\vec{y}} + A_3[\delta_{\vec{x},\vec{x}_0} + \delta_{\vec{y},\vec{x}_0}] + A_4\delta_{\vec{x},\vec{x}_0}\delta_{\vec{y},\vec{x}_0}\,, \tag{C16}$$

$$B(\vec{x},\vec{y}) \;=\; B_1 + B_2\delta_{\vec{x},\vec{y}} + B_3[\delta_{\vec{x},\vec{x}_0} + \delta_{\vec{y},\vec{x}_0}] + B_4\delta_{\vec{x},\vec{x}_0}\delta_{\vec{y},\vec{x}_0}\,, \tag{C17}$$

where (defining $z \equiv 1 - \gamma(1-2^{-n})(1-\varphi^k)$)

$$A_1 \;=\; -\frac{1}{z}\,k(k-1)\gamma(1-\varphi^{k-2}) + \frac{1}{z^2}k^2\gamma^2\,\frac{1}{2^n}(1-\varphi^{k-1})^2\,, \tag{C18}$$

$$A_2 \;=\; -\frac{1}{z}\,k(k-1)\gamma\varphi^{k-2}\,, \tag{C19}$$

$$A_3 \;=\; \frac{1}{z^2}\,k^2\gamma^2\,\frac{1}{2^n}(1-\varphi^{k-1})\varphi^{k-1}\,, \tag{C20}$$

$$A_4 \;=\; \frac{1}{z^2}k^2\gamma^2\varphi^{2(k-1)}\,, \tag{C21}$$

and (defining $\Lambda_0 = (e^\omega - 1 + 2^n)^{-1}$ and $\Lambda = 1 - 2^{n\Lambda_0}$)

$$B_1 \;=\; -2^n\Lambda_0^2\,, \qquad B_2 = \Lambda_0\,, \tag{C22}$$

$$B_3 \;=\; -\Lambda\Lambda_0\,, \qquad B_4 = \Lambda(1-\Lambda)\,. \tag{C23}$$

The quadratic form in Eq. (C14) can be diagonalized both in position space (by Fourier transform) and in replica space (all the eigenvectors have the form $\zeta(\vec{x}) = \zeta_0 \delta_{\vec{x},\vec{x}_0} + \zeta_1$). One can therefore perform the gaussian integral, and let $n \to 0$. Putting this contribution together with the action correction, cf. Eq. (C12), we finally get the entropy correction

$$\phi^{(1)}(\gamma) = -\binom{k}{2}\gamma\varphi^{k-2}(1-\varphi^2) - \int_{-\infty}^{+\infty}\left\{\log(1-a\,w(q)) + \frac{b\,w(q)}{1-a\,w(q)}\right\}\frac{\mathrm{d}q}{4\pi}\,, \tag{C24}$$

where

$$a \;=\; k(k-1)\gamma\varphi^{k-2}e^{-\omega}\,, \tag{C25}$$

$$b \;=\; -k(k-1)\gamma\varphi^{k-2}e^{-\omega}(1-e^{-\omega}) + k^2\gamma^2\varphi^{2(k-2)}e^{-\omega}\,. \tag{C26}$$

---

[1] T. R. Kirkpatrick and P. G. Wolynes, Phys. Rev. B **36** (1987) 8552

[2] T. R. Kirkpatrick, D. Thirumalai, and P. G. Wolynes, Phys. Rev. A **40** (1989) 1045

[3] J.-P. Bouchaud, L. F. Cugliandolo, J. Kurchan and M. Mézard, "Out of equilibrium dynamics in spin-glasses and other glassy system", in *Spin Glasses and Random Fields*, A. P. Young ed., (World Scientific, Singapore, 1997)

[4] Biroli G, Monasson R and Weigt M (2000) *Eur. Phys. J. B* **14**, 551-568.

[5] Mézard M, Parisi G and Zecchina R (2002) *Science* **297**, 812-815.

[6] F. Krzakala, A. Montanari, F. Ricci-Tersenghi, G. Semerjian, and L. Zdeborova, 'Gibbs States and the Set of Solutions of Random Constraint Satisfaction Problems,' `arXiv:cond-mat/0612365`, and Proc. Natl. Acad. Sciences, in press.

[7] P. G. Wolynes, Jour. Res. NIST **102** (1997) 187,

[8] X. Xia, P. G. Wolynes, Proc. Nat. Acad. Sci. **97**, (2000) 2990

[9] X. Xia, P. G. Wolynes, Phys. Rev. Lett **86** (2001) 5526

[10] J.-P. Bouchaud and G. Biroli, J. Chem. Phys. **121** (2004) 7347

[11] A. Montanari and G. Semerjian, J. Stat. Phys. **125**, 23 (2006).

[12] S. Franz and A. Montanari, J. Phys. A **40** (2007), F251-F257

[13] N. Creignou and H. Daudé, Discrete Appl. Math. **96-97** 41 (1999).

[14] F. Ricci-Tersenghi, M. Weigt and R. Zecchina, Phys. Rev. E **63**, 026702 (2001).

[15] M. Mézard, F. Ricci-Tersenghi and R. Zecchina, J. Stat. Phys. **111**, 505 (2003).

[16] S. Cocco, O. Dubois, J. Mandler and R. Monasson, Phys. Rev. Lett. **90**, 047205 (2003).

[17] A. Montanari and G. Semerjian, Phys. Rev. Lett. **94**, 247201 (2005).

[18] A. Montanari and G. Semerjian, J. Stat. Phys. 124, 103 (2006)

[19] J. M. Schwarz and A. A. Middleton, Phys. Rev. **E 70** (2004) 035103 (R)

[20] S. Franz and F. L. Toninelli, Phys. Rev. Lett. **92** (2004) 030602

[21] S. Franz and F. L. Toninelli, J. Phys. A: Math. Gen. **37** (2004) 7433

[22] S. Franz and F. L. Toninelli, J. Stat. Mech. (2005) P01008

[23] S. Franz and G. Parisi, Europhys. Lett. **75** (2006) 385-391

[24] N. Alon and J. Spencer, *The Probabilistic Method*, Wiley, New York, 1992.

[25] R. Monasson, J. Phys. A **31** (1998) 513-529